**Introdução**

A inteligência artificial tem sido cada vez mais empregada como impulsionadora do desenvolvimento tecnológico, agilizando processos e simplificando desde tarefas básicas até as mais complexas. Dentro das diversas aplicações da IA, o algoritmo K-means é uma técnica de aprendizagem não supervisionada bastante consolidada que não necessita de dados rotulados nos parâmetros de entrada do modelo. Tal algoritmo tenta encontrar padrões em grandes conjuntos de dados, a partir de um procedimento de clusterização, em que cada ponto de dados é atribuído ao cluster cujo centroide está mais próximo. Essa técnica é bastante difundida e consolidada dentro da inteligência artificial devido à sua simplicidade, eficiência e aplicabilidade.

Com isso, uma possível aplicação desse algoritmo está voltada ao reconhecimento de eletrofácies a partir de dados de perfilagem. Diante disso, sabe-se que a litologia fornece informações sobre a composição mineralógica, estrutura porosa e propriedades físicas das rochas, o que influencia diretamente na capacidade de armazenamento e na fluidez do petróleo. Compreender estes aspectos ajuda na identificação de reservatórios de petróleo, na determinação das técnicas de extração mais adequadas e na previsão da produtividade dos poços. Dessa forma, para obtenção de modelos precisos de reservatórios, o reconhecimento litológico é uma fase essencial. Para o trabalho, foi utilizado o algoritmo K-Means para realizar a classificação de eletrofácies a partir de dados de perfis geofísicos de poços onshore da Bacia Potiguar, a fim de identificar padrões complexos e fazer previsões sobre a litologia das rochas subsuperficiais.

1. **Desenvolvimento**
   1. **Perfilagem de Poços**

Na geologia do petróleo, uma maneira de identificar as propriedades das rochas em subsuperfície é através dos perfis geofísicos de poços, que são responsáveis por medir indiretamente as propriedades das rochas. Esses perfis geofísicos de poços são capazes de identificar a porosidade, densidade, resistividade e radioatividade no reservatório. (TORRES; BATEZELLI, 2020). Partindo dessas medições, os geólogos e geofísicos são aptos a encontrar padrões em diferentes conjuntos rochosos. Tais padrões são denominados eletrofácies (DE SOUSA; CHAVES-SGARBI, 2019). De acordo com Serra e Abbott (1982), eletrofácies podem ser definidas como um conjunto de respostas de perfis elétricos que caracterizam fácies litológicas e sedimentares, permitindo distingui-las das outras. Dessa forma, na interpretação de perfis, as eletrofácies são identificadas como classes numericamente definidas, podendo ser analisadas de várias maneiras: manualmente, combinando diferentes perfis; através de gráficos de dispersão (crossplots); usando gráficos de histogramas; ou por meio de métodos de análise estatística (RIDER, 2002) A partir disso, segundo Kumar e Kishore (2006), comumente as eletrofácies têm sido identificadas manualmente com o auxílio de técnicas gráficas, como o cruzamento de dados de registros de poço. No entanto, as pessoas frequentemente estão sujeitas a cometer erros durante análises ou, possivelmente, ao tentar estabelecer relações entre múltiplas características. Conforme menciona Torres (2022), os problemas envolvem não-padronização nas classificações de eletrofácies, isto é, a possibilidade da obtenção de diferentes padrões de classificação de acordo com cada geólogo, e o número muito grande de registros de poços, tornando a classificação manual cara e longa. Com base nisso, diversos algoritmos de inteligência artificial têm sido implementados visando a automatização de reconhecimento de eletrofácies (KOLKATA, 2006). Por fim, destaca-se a grande importância do reconhecimento de eletrofácies para os estudos de avaliação de perfis de poço e caracterização de litofácies (ROSA, 2006), ou também na caracterização de fácies reservatórios (GROU, 2015).

O presente estudo foi desenvolvido a partir de dados de treinamento e teste. Os dados de treinamento correspondem ao perfil composto do poço 4-BRSA-814D-RN, localizado no bloco BT-POT-8, no campo GOES, na Bacia Potiguar onshore. Já os dados de teste correspondem ao perfil composto do poço 4-BRSA-937D-RN, localizado em um bloco não informado, no campo Jaçanã, na Bacia Potiguar onshore. Por último, é importante destacar a essencialidade dos poços possuírem formações correlatas para que o estudo assegurasse sua validade nos resultados obtidos.

* 1. **Aprendizado Não Supervisionado e Algoritmo K-Means**

Nos últimos anos, a indústria de óleo e gás tem agregado cada vez mais os algoritmos de inteligência artificial nos mais diversos setores como também no desenvolvimento de pesquisas visando o desenvolvimento tecnológico da indústria (KURODA et al., 2012; SFIDARI et al., 2012). A grande quantidade de dados gerados tem demandado metodologias mais robustas de análise para a geração de insights. Nesse sentido, a compreensão abrangente das distribuições do reservatório e suas características de desempenho de produção necessita do processamento e análise de uma grande quantidade de dados (EUZEN et al., 2010). No conjunto da inteligência artificial existem dois grandes grupos de algoritmos que são bastante utilizados, *Aprendizagem Supervisionada* (Kotsiantis, 2007) e *Aprendizagem Não supervisionada* (Jain, 2010).

No algoritmo de aprendizagem supervisionada, os dados de treinamento são compostos por partes de entrada e saída conhecidos, denominados rótulos. Já no algoritmo de aprendizagem não supervisionada, os dados de treinamento não possuem rótulos ou saídas conhecidas.

A presente pesquisa foi inteiramente direcionada para aprendizagem não supervisionada, haja vista que em um contexto de aprendizagem supervisionada necessitaria de um esforço manual extra para examinar os registros de poços e atribuir rótulos aos dados de entrada, dessa forma seria bastante custoso devido ao grande conjunto e a complexidade dos dados. Diante disso, a aprendizagem não supervisionada além de não solicitar a rotulação de dados permite uma exploração da estrutura intrínseca dos dados, identificando padrões ou clusters que dificilmente seriam obtidos com a presença de rótulos externos.

Dessa forma, o algoritmo não supervisionado escolhido para a realização da pesquisa foi o k-means, haja vista sua eficácia e sucesso obtido em outros estudos de aplicação descritos por (AUTORES). Apesar disso, é importante destacar que mesmo o algoritmo k-means apresentando um alto potencial, após o processo de classificação, o conhecimento humano é necessário para interpretar a classificação feita pelo modelo e na aprendizagem supervisionada é diferente, pois inicialmente precisamos ensinar o modelo, mas após o processo de ensino o conjunto de dados classificado não precisa ser interpretado.

A clusterização k-means é um método comumente usado para particionar automaticamente um conjunto de dados em k grupos (WAGSTAFF et al., 2001). A clusterização de dados, também conhecida como análise de agrupamento, tenta descobrir o(s) agrupamento(s) natural(is) de um conjunto de padrões, pontos ou objetos (JAIN, 2010). O objetivo da análise de agrupamento é classificar um conjunto de dados em grupos que sejam internamente homogêneos e externamente isolados com base em uma medida de similaridade ou dissimilaridade entre os grupos (KIAEI, 2015).

Dado um conjunto de 𝑛 pontos de dados no espaço real 𝑑-dimensional, , e um número inteiro 𝑘, o problema consiste em determinar um conjunto de 𝑘 pontos em , chamados centros, de modo a minimizar a distância média ao quadrado de cada ponto de dados ao seu centro mais próximo." (KANUNGO et al., 2002). A Figura X apresenta um diagrama simplificado do funcionamento do algoritmo k-means.

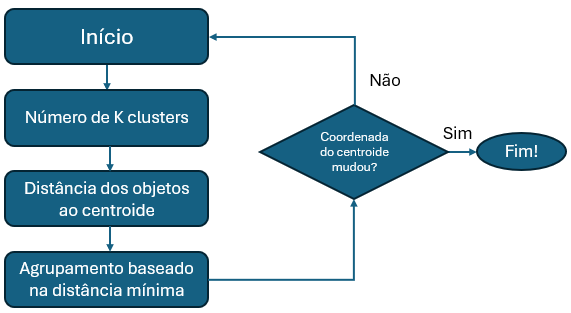


Figura X. Algoritmo k-means.

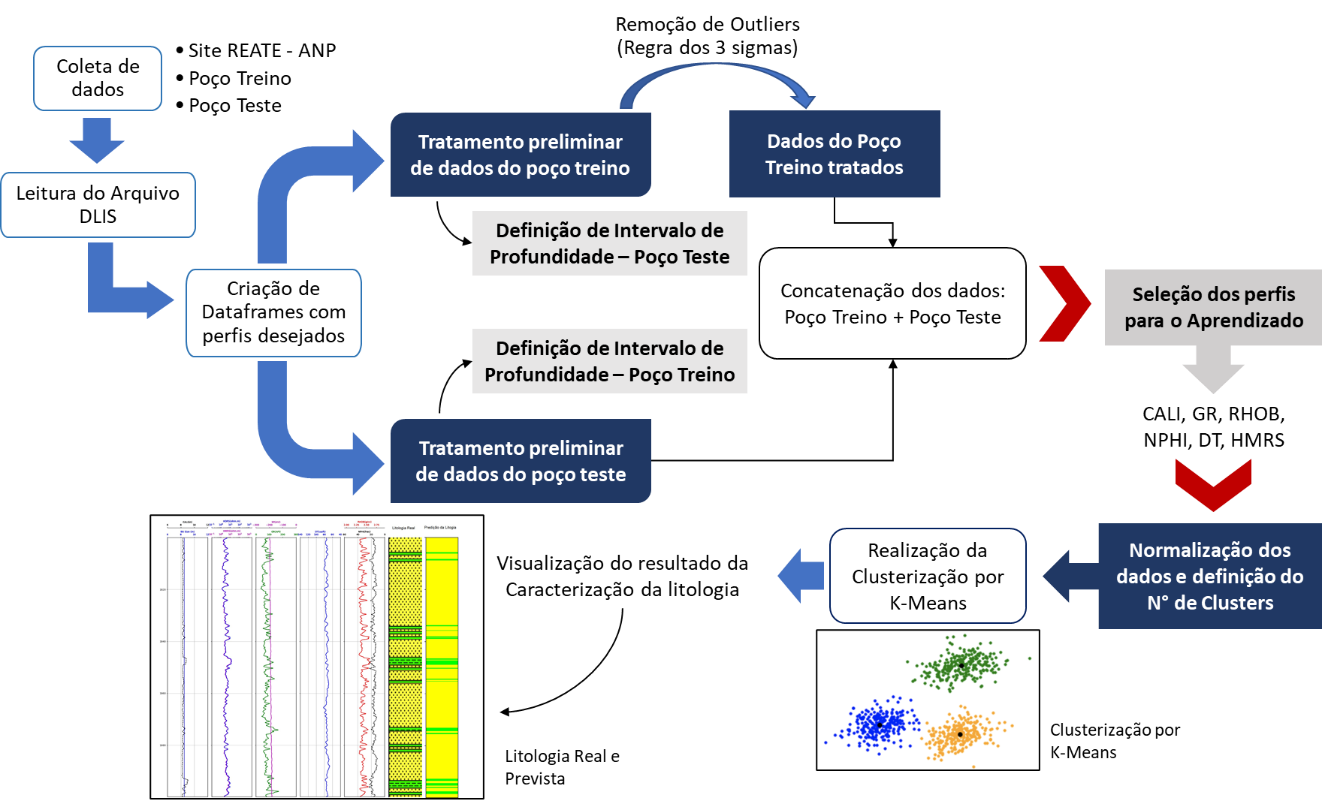
O algoritmo foi implementado a partir da biblioteca *Sckit-learn* presente na linguagem de programação Python.

Além disso, foi utilizado o método do cotovelo (Elbow Method) para encontrar o número ideal de clusters. Esse método é baseado na variação de inércia - também conhecida como soma das distâncias quadráticas entre cada ponto e o centroide do cluster ao qual ele foi atribuído - em relação ao número de clusters, em outras palavras, quanto menor a inércia, mais compactos são os clusters O gráfico é gerado a partir da correlação entre o número de clusters (eixo X) e a soma dos erros quadrados (eixo Y). O ponto de cotovelo no gráfico será o ponto onde a inércia começa a decrescer em uma taxa mais lenta, formando um padrão semelhante a um cotovelo.

1. **Metodologia**

A estruturação do projeto parte da execução de seguidas etapas fundamentais para obtenção de resultados adequados e confiáveis. A figura 1 abaixo mostra o fluxograma do processo empregado desde a coleta e tratamento dos dados até a execução do algoritmo de clusterização para obter a predição das eletrofácies de um poço em determinado intervalo de profundidade.

**Figura 1 –** Fluxograma do processo de predição das eletrofácies



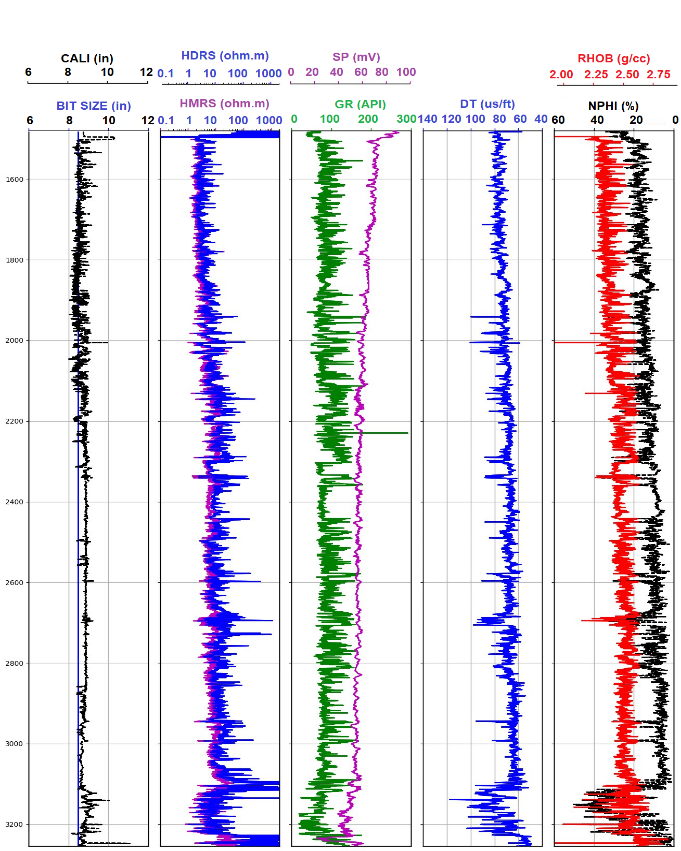
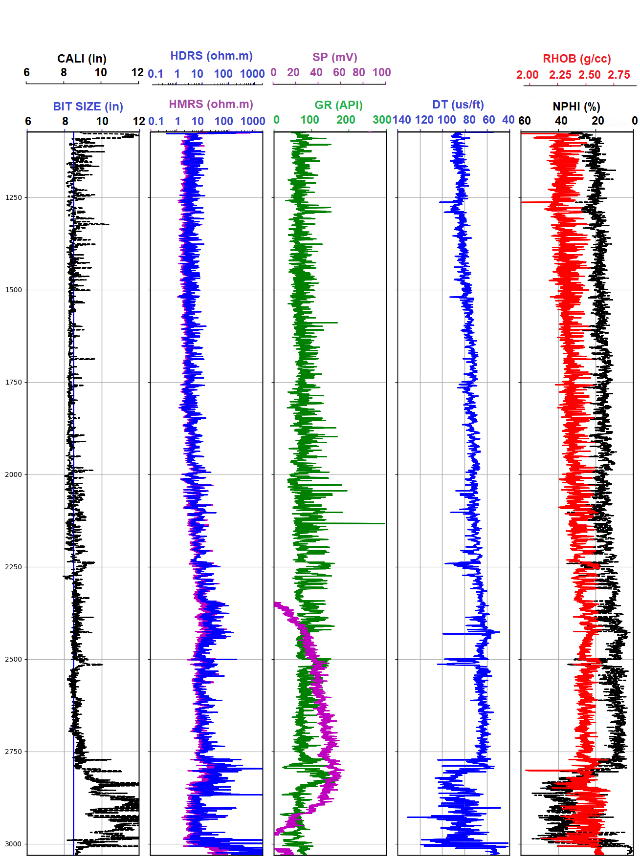
Fonte: Autor (2024)

* 1. **Coleta, Leitura e Tratamento de dados**

Os dados utilizados no projeto foram retirados do site REATE, da ANP, o qual reúne informações com dados técnicos de bacias sedimentares terrestres brasileiras e possui vários dados de poços, entre eles os de perfilagem. Para o trabalho, foram definidos dois poços da Bacia Potiguar para realizar a predição de eletrofácies, poço 4-BRSA-814D-RN e 4-BRSA-937D-RN. Os arquivos dos poços com os dados dos perfis estão disponíveis em formato DLIS, comumente utilizados para representar uma variedade de dados geofísicos complexos.

A leitura dos dados é feita por meio da utilização da biblioteca **dlisio** em Python. A partir da leitura do arquivo, os dados são modelados e tratados de forma eficiente. O intervalo de profundidade é selecionado, assim como os perfis adequados, mostrados na figura abaixo, para treinar a máquina: Cáliper, densidade, neutrão, sônico, gamma-ray e resistividade.

**Figura 2** – Perfis dos poços 4-BRSA-814D-RN (a) e 4-BRSA-937D-RN (b).

 (a)  (b)

Fonte: Autor (2024)

Depois desta etapa, os dados de treino (Poço 4-BRSA-814D-RN) são analisados e os outliers (valores que fogem drasticamente da normalidade) do perfil de Cáliper (Diâmetro do Poço) são identificados pelo método dos 3 sigmas (Valores abaixo e acima de 3 desvios padrões da média) e excluídos para utilizar no modelo, a fim de realizar o controle de qualidade dos dados.

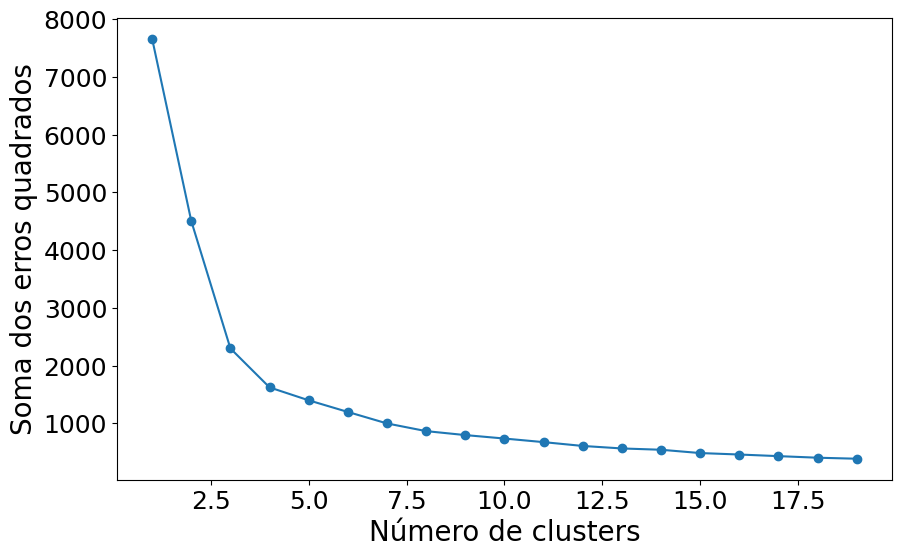
* 1. **Clusterização dos dados**

Posteriormente a etapa de tratamento dos dados de treino, os dados de teste, os quais são em menor quantidade (Proporção de 80% de dados de treino e 20% de dados para teste), são determinados e modelados, sem retirada dos outliers, e, então, concatenados junto com os dados de treino, a fim de promover a clusterização dos valores a partir desses dados disponibilizados. Para o projeto, foi selecionado o intervalo de 2.400 a 2.800 metros do poço treino e utilizada a profundidade de 1.800 a 1.900 metros do poço teste por apresentar condições ideais.

Entretanto, antes de aplicar o algoritmo de agrupamento K-Means, é necessário que os dados sejam padronizados ou normalizados. Esta é uma etapa importante no pré-processamento de dados e visa transformar as características (variáveis) para que elas tenham média zero e desvio padrão igual a um, uma vez que os dados (Os perfis) apresentam diferentes escalas.

Após esta etapa, faz-se necessário definir o número de clusters para a predição das eletrofácies, em que cada cluster, ou grupo, representa uma eletrofácie. Para isso, foi utilizado o método da Inércia para descobrir o melhor número de clusters, onde essa técnica calcula a soma das distâncias quadradas dos pontos aos centróides. Quando esses valores são inseridos graficamente, a definição do número adequado de clusters se dá pelo ponto de cotovelo, momento pelo qual ocorre a inflexão da curva e é possível fazer a identificação visual. A figura abaixo mostra o gráfico obtido com os dados do projeto para a definição do número de agrupamentos.

**Figura 3** - Determinação do número de clusters pelo método do cotovelo.



Fonte: Autor (2024)

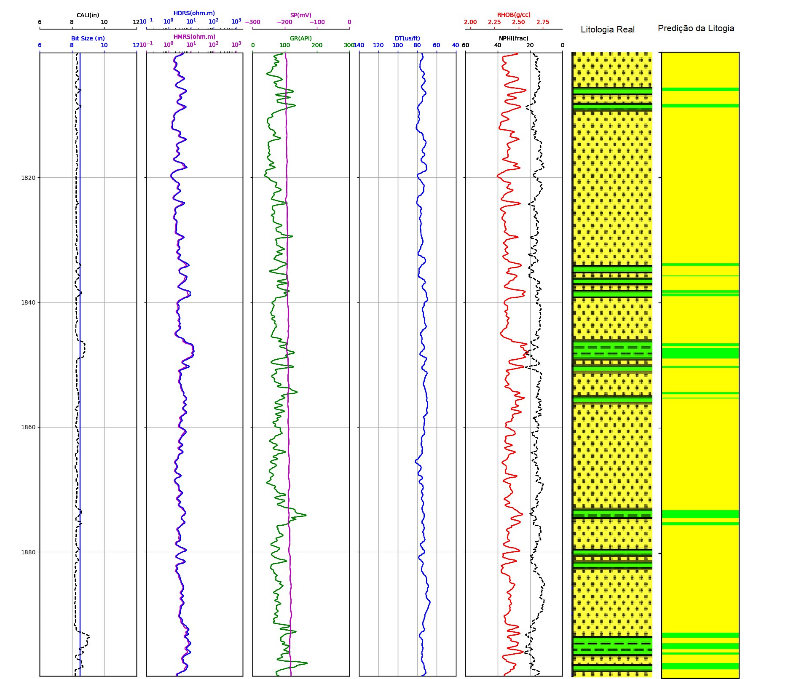
Pelo gráfico, é possível observar que 3 ou 4 clusters já são adequados para o caso, uma vez que um número maior de clusters pode trazer erros de identificação, além de que o perfil composto do poço aponta apenas 2 litologias predominantes, arenito e folhelho.

Após a definição do número adequado de clusters, aplica-se o algoritmo de clusterização K-Means em cima dos dados e os resultados são, então, obtidos e analisados.

1. **Resultados e Discussões**

Os resultados obtidos neste estudo revelaram uma semelhança satisfatória entre a litologia real do poço 4-BRSA-937D-RN e a litologia predita para o mesmo a partir do aprendizado de máquina para o intervalo de 1.800 a 1.900 metros, como mostra a figura 4. É possível observar espaços bem definidos que o algoritmo foi eficiente em representar, entretanto apresenta outros que não foi capaz de identificar ou classificar no cluster correto.

Figura 2. Comparação entre a litologia real e a predita.



Fonte: Autor (2024)

Na tentativa de quantificar a semelhança entre as eletrofácies predita e real, foi utilizada uma técnica de comparação de pixels por meio do Python e utilizando a biblioteca OpenCV, que oferece métodos para comparar imagens. Uma das funções disponíveis é a **cv2.absdiff()**, a qual calcula a diferença absoluta entre as duas imagens e, em seguida, calcula a porcentagem de semelhança com base na diferença encontrada. Para o caso do projeto, a função calculou a semelhança entre as imagens e forneceu o valor de **95,19%** de correspondência.